UNIVERSIDAD TÉCNICA DE MANABÍ
Fundada en 1952

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Covid-19 en Ecuador: Aplicación de minería de datos

Covid-19 en Ecuador: Aplicación de minería de datos

Autores

✉ * *Juan Carlos Zambrano*

✉ *Patricia Quiroz-Palma*

✉ *Alex Santamaría-Philco*

✉ *Willian Zamora*

[1]Facultad de Ciencias Informáticas, Universidad Laica Eloy Alfaro de Manabí, Manta, Ecuador.

* Autor para correspondencia

Comó citar el artículo: Zambrano, J. C., Quiroz-Palma, P., Santamaría-Philco, A., & Zamora, W. (2022). Covid-19 en Ecuador: Aplicación de minería de datos. *Informática y Sistemas: Revista de Tecnologías de la Informática y las Comunicaciones, 6*(1), 12-23. DOI: https://doi.org/10.33936/isrtic.v6i1.4366

Enviado: 31/01/2022
Aceptado: 09/05/2022
Publicado: 29/05/2022

Resumen

El COVID-19 se introdujo rápidamente como una pandemia global, la cual necesita ser tratada con respuesta inmediatas e integradas a todos los sistemas nacionales que requieran de ellas. Con la llegada del COVID-19 el mundo vio la necesidad de respuestas oportunas y el intercambio de datos sobre ésta y futuras pandemias globales de rápida propagación. Este estudio se enfoca en predecir la incidencia del COVID-19 en Ecuador. Se realizó Minería de Datos de los registros proporcionados por instituciones públicas del estado ecuatoriano con información oficial y actualizada del COVID-19 en el Ecuador. Se experimento con modelos de regresión y memoria a largo plazo obteniendo como resultado el modelo óptimo para estimar el número de casos positivos de COVID-19. Para el modelo matemático se hizo uso del error cuadrático medio como métrica del rendimiento. Del análisis de los datos sobre el COVID-19 en Ecuador el modelo de regresión lineal predijo la incidencia con un error cuadrático medio de 0.54. siendo los factores más efectivos la incidencia de días anteriores y el número de población de cada una de las provincias afectadas.

Palabras clave: Covid-19; Minería de datos; Metodología KDD; Algoritmos series-temporales.

Abstract

COVID-19 was quickly introduced as a global pandemic, which needs to be addressed with immediate and integrated responses to all national systems that require them. With the advent of COVID-19 the world saw the need for timely responses and data sharing on this and future rapidly spreading global pandemics. This study focuses on predicting the incidence of COVID-19 in Ecuador. Data mining was performed on records provided by public institutions of the Ecuadorian state with official and updated information on COVID-19 in Ecuador. We experimented with regression and long-term memory models, obtaining as a result the optimal model to estimate the number of positive cases of COVID-19. For the mathematical model, the mean square error was used as a performance metric. From the analysis of the data on COVID-19 in Ecuador, the linear regression model predicted the incidence with a mean square error of 0.54, the most effective factors being the incidence of previous days and the population in each of the affected provinces.

Keywords: Covid-19; Data mining, KDD methodology, Time-series algorithms.

https://revistas.utm.edu.ec/index.php/informaticaysistemas/index

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

1. Introducción

El mundo entero ha sido afectado de los efectos producidos por la pandemia provocada por coronavirus (COVID-19). La mayoría de los países se registran miles de infectados, bajas humanas, pacientes con secuelas y gran afectación económica. El coronavirus surgió en Wuhan, China y se extendió por todo el mundo. El análisis genómico reveló que el SARS-CoV-2 está relacionado filogenéticamente con los virus de murciélagos, similares al síndrome respiratorio agudo severo (similar al SARS), determinándose que estos mamíferos podrían ser el posible reservorio primario. Se desconoce la fuente intermedia de origen y transmisión a humanos, sin embargo, la rápida transferencia de humano a humano se ha confirmado ampliamente. (Muhammad, Suliman, Abeer, Nadia, & Rabeea, 2020).

En Ecuador se han reportado alrededor de 859 mil casos para el mes de marzo del 2022 desde el inicio de la emergencia sanitaria, según el informe emitido por el Ministerio de Salud Pública. Siendo Pichincha la provincia más afectada con 64469 casos en total representando el 36.99% de los casos registrados a escala nacional. En tanto, hay 35.421 fallecidos en el contexto de la pandemia en Ecuador. En Ecuador según los datos de (MSP, 2022) se han aplicado más de 30 millones de dosis de la vacuna contra el virus. Los resultados de esta se evidenciaron al registrar un número menor de casos confirmados por Covid-19 en comparación a fechas anteriores a la vacuna. Sin embargo, entre la segunda y tercera semana del mes de enero del año 2022 se registraron más de 100 mil casos, mientras que durante la ultima semana del 2021 se registraron 9092 casos, evidenciándose un gran incremento en de nuevos casos. Estos datos representan una valiosa fuente de información para estudiar el compartimiento del virus dentro del contexto ecuatoriano. Sin embargo, son pocos los estudios que apliquen modelos predictivos que generen pronósticos y que estos a su vez respalden a la toma de decisiones de los ministerios y organismos para salvaguardar la seguridad ciudadana.

En la India, (ubicar el espacio temporal) que al alcanzar las cifras de 536 casos y 11 muertos tomo la decisión de ejecutar un bloqueo nacional y rápidamente llamo al campo a las ciencias de datos para reunir información e implementar estrategias de predicción. Una de ellas fue una extensión bayesiana del modelo Susceptible-Infected-Removed (eSIR) diseñado para el pronóstico de intervenciones para estudiar el impacto a corto y largo plazo de un bloqueo inicial de 21 días en el número total de infecciones por COVID-19 (Debashree, Maxwell, Rupam, & Lili, 2020). En Ecuador se realizó un modelo predictivo de los casos de contagio por Covid-19 para la provincia de Loja presentado por los autores (Salcedo & Salcedo, 2021), donde se aplica método numérico de diferencias divididas, método logístico y método de mínimos cuadrados para predecir el nivel de contagio. Para ello se hizo uso de los datos proporcionados por el Ministerio de Salud pública del Ecuador en un periodo de 399 días.

En este trabajo se presenta el proceso de minería de datos para predecir la incidencia del Covid-19 en el Ecuador, para se realizaron cada una de las etapas de la metodología Knowledge Discover Database (KDD). Los datos analizados durante un año nos permitieron predecir la proyección de incidencia de contagios en el Ecuador y sus provincias. En la sección 2, se describen los materiales y métodos usados en la solución propuesta mediante procesos de minería de datos. En la sección 3, se describen los resultados obtenidos de las predicciones; en la sección 4, se describen la discusión y en la sección 5 se describen las conclusiones y trabajos futuros de este trabajo de investigación.

2. Materiales y Métodos

Los investigadores de todo el mundo están utilizando el aprendizaje automático (Machine Learning) para desarrollar modelos que simulan y predicen la propagación del virus en un intento por identificar patrones que puedan revelar las debilidades y los peligros de esta pandemia. Robert H (Jason & Robert, 2020) expresa lo siguiente: “Taiwán aprovechó su base de datos de seguros de salud y la integró con su inmigración y base de datos aduanera para iniciar la creación de big data para análisis; generó alertas en tiempo real durante una visita clínica basada en el historial de viajes y síntomas para ayudar a la identificación del caso”

Tomando en cuenta lo expuesto y considerando el estado actual de la crisis sanitaria en Ecuador es necesario preparar un modelo de minería de datos que prestará los siguientes beneficios:

• Se recopilaron los registros diarios del Ministerio de Salud Pública del Ecuador y se lo modelara en un conjunto de datos el mismo que será accesible para que cualquier investigador o entidad a quien interese haga uso de los datos.

• Los modelos de regresión lineal indicarán las provincias presentarán mayor incidencia, lo que apoyaría a la toma de decisiones de las prefecturas y municipios competentes.

La investigación, tiene un enfoque descriptivo e interpretativo, con información de tipo cuantitativo, que busca predecir el número de contagios por COVID-19 en Ecuador. Para ello se consideran los datos proporcionados por el ministerio de Salud Pública del Ecuador, que posteriormente son recopilados entre el 13 de marzo del 2020 hasta 28 de marzo del 2022 el en un conjunto de datos.

2.1. Métodos de Investigación

Los métodos de investigación que se emplearon para predecir

e-ISNN 2550-6730 | Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022 | DOI: 10.33936/isrtic.v6i1.4366

la incidencia del COVID-19 en el Ecuador se describen a continuación:

• Método Analítico: Consiste en la desmembración de los elementos que componen la totalidad de las técnicas de minería de datos para la predicción, también instaura las relaciones causa, efecto y la naturaleza del problema identificado.

• Método Sintético: Con este método se pretende reconstruir en un todo a partir los elementos distinguidos por el análisis de minería de datos para estudiarlos y conseguir un extracto minucioso para la predicción de la incidencia del virus con técnicas de minería de datos.

• Método Deductivo: Se emplea este método para analizar la minería de datos, yendo de lo general a lo particular para definir conclusiones que ayudan a la predicción de la incidencia del COVID-19 utilizando registros históricos de los casos que se presentaron y se siguen presentando en el país, además de los factores que tengan relación con el incremento de contagios.

• Método Inductivo: Este método se operó realizando observaciones de la teoría y trabajo de otros investigadores entre estos el modelo predictivo para la incidencia del COVID-19 en Iran mediante el análisis de datos de Google Trends (Ayyoubzadeh, Zahedi, & R, 2020), el modelo predictivo de la salud de pacientes por COVID-19 aplicando algoritmos de bosques aleatorios (Celestine, Ali, Atharva, & R, 2020) en ellos se utilizaron técnicas de minería de datos para predecir la incidencia del COVID-19, mediante el razonamiento a partir de premisas particulares para sustentar conclusiones.

2.2 Herramientas de Recolección de datos

Las herramientas de recolección de datos empleadas para esta investigación son las siguientes.

El análisis documental: Se recopila información de diversas fuentes como lo son, reportes del MSP, COE Nacional, libros, sitios web, artículos científicos, etc., que tenga un enfoque en el problema identificado. Esta exploración de contextos, además de usarse para el desarrollo del marco teórico conceptual referente a minería de datos, incluida las fases, técnicas, algoritmos y demás; también tuvo su lugar en la extracción de los datos con relevancia para formar la conjunto de datos fundamental en el entrenamiento de los algoritmos de predicción empleados.

La observación: El investigador hace uso de su razonamiento luego de observar y percibir los aconteceres locales y nacionales que se difunden por diversos medios y que tienen relación con el comportamiento de la pandemia a causa del COVID-19, permitiéndole identificar los objetos de estudios como patrones de propagación, incrementos en las tasas de contagio, aplicación de medidas de bioseguridad, etc., estos factores se correlacionan con la incidencia del virus.

La encuesta: Como herramienta adicional de recolección de datos para la investigación se empleó la encuesta, la misma que es apoyada por cuestionarios, como instrumento de recolección de la información. Estos cuestionarios fueron aplicados a través de la herramienta Google Form y repartidos en grupos de redes sociales, Whatsapp y Facebook cuyos grupos están integrados por miembros de la comunidad estudiantil de las Facultades de Ciencias Informáticas, Ciencias Médicas y Enfermería de la Universidad Laica Eloy Alfaro de Manabí, además un pequeño número de formularios fueron extendidos hasta personas que laboran en primera línea en lucha contra la pandemia de COVID-19. Se optó por tomar una muestra de este universo puesto que varias de las preguntas incluidas en los formularios son de campo y requerían de un poco de conocimiento técnico para ser respondidas de manera confiable.

2.3 Fuentes de información de datos

Fuentes primarias: Para el desarrollo y propósito de esta investigación se optó por tomar el análisis documental como herramienta principal de información, siendo las fuentes más importantes los boletines epidemiológicos emitidos diariamente por el MSP , y COE Nacional esta información se encuentra documentada en los repositorios oficiales de las instituciones antes mencionadas, de donde se extrajeron datos de gran valor para el estudio de la predicción de la incidencia del virus en el país, como lo son el número de nuevos casos de cada provincia, dadas de altas, fallecimientos, vacante hospitalaria, género y rango de edad de los pacientes intervenidos.

Fuentes Secundarias: Como fuentes secundarias de información se designan los resultados obtenidos en las encuestas realizadas al muestreo definido de la población, que en conjunto con una observación estructurada realizada por parte del investigador hacia los aconteceres nacionales relacionados a la pandemia que enfrenta el Ecuador, se logra identificar coincidencias en algunas de las variables consideradas como factores que intervienen en la incidencia del COVID-19 en el país.

2.4 Proceso de Minería de Datos

El proceso Knowledge Discover Database (KDD), tal como se presenta en (Fayyad, Piatetsky, & Smyth, 1996), es el proceso de usar métodos de Data Mining (DM) para extraer conocimiento, utilizando una base de datos junto con cualquier preprocesamiento, submuestreo y transformación requeridos de la base de datos. Se consideran cinco etapas como se muestra en la Figura 1.

1. Selección. - Esta etapa consiste en crear un conjunto de datos objetivo, o enfocarse en un subconjunto de variables o muestras de datos, en las que se debe realizar el descubrimiento.

2. Procesamiento Previo. - Esta etapa consiste en la limpieza y preprocesamiento de datos de destino para obtener datos consistentes

3. Transformación. - Esta etapa consiste en la transformación de los datos usando métodos de reducción de dimensionalidad o transformación.

4. Minería de Datos. - Esta etapa consiste en la búsqueda de patrones de interés en una forma representacional particular,

e-ISNN 2550-6730 | **Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022** | **DOI: 10.33936/isrtic.v6i1.4366**

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

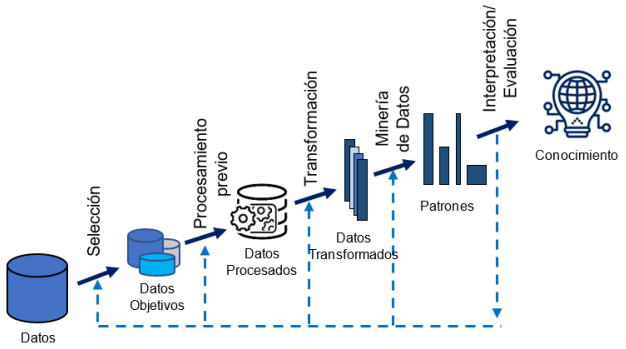


Figura 1. Etapas en el procedimiento KDD.

dependiendo del objetivo de minería de datos (por lo general, la predicción).

5. Interpretación / Evaluación. - Esta etapa consiste en la interpretación y evaluación de los patrones.

2.4.1. Selección de Datos

Para la obtención de la información se precedió a realizar el análisis de las infografías emitidas por el Servicio Nacional de Gestión de Riesgos y Emergencias desde su repositorio digital (COE, 2022). A través del cual se obtuvieron trescientos ochenta y dos registros en un periodo de 365 días. Para los mismos se definió una estructura adaptada a un conjunto de datos tabulados para su procesamiento.

La siguiente fase de esta etapa correspondió a la creación del conjunto de datos siendo la fuente de datos los informes por casos de COVID-19 del COE Nacional, dichos informes se emitieron en documentos digitales con extensión de archivo '.pdf' que representa una fuente de datos no procesados. Además de esto, todos los informes presentan un formato general que comprende su contenido, esto incluye: textos, tablas, imágenes, gráficos estadísticos, marcas de agua etc., Las condiciones de los documentos dificultó la extracción automática de los datos. Por este motivo, se planteó un método alternativo para la extracción de los datos, el cual implica: la extracción por captura de imagen del contenido relevante de los informes, aplicación de filtro de color para minimizar las marcas de agua, la aplicación de una herramienta de reconocimiento óptico de caracteres y libros de Excel. Para la comprensión de esta fase se detalla el proceso en la Figura 2 con la descripción de cada uno de los pasos.

2.4.2. Procesamiento previo

Después del análisis y selección de Python con sus librerías para el análisis de datos como herramienta de minería de datos, se procede a continuar con esta etapa de la metodología. De este modo se realiza la carga de información de los conjuntos de datos "Covid-19EC_dataset.csv" y "Datos provincia.csv" los mismo que fueron almacenados en variables de nombre "df1" y "df2" respectivamente, esto se logró mediante mediante el paquete de Pyton Pandas por lo que se apoya en el uso del módulo 'pandas. read_csv', que proporciona estructuras de datos similares a los marcos de datoss (ver Figura 3).

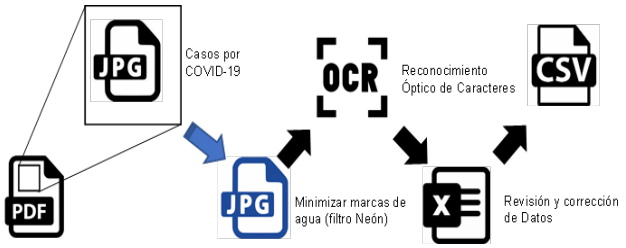


Figura 2. Proceso de extracción de datos.

Así los datos pasan de ser visualizados como se muestra en la Figura 4, en su estructura original del conjunto de datos en formato CSV delimitado por comas, a su adaptación dentro de un marcos de datos proporcionado por la librería Pandas como se observa en las Figuras 5 y 6.

Después de cargar el conjunto de datos y, siguiendo las fases de la metodología se procedió con el preprocesamiento de los datos el cual consiste en dos etapas: 'limpieza' y 'procesamiento'.

Se procedió a concatenar los conjuntos de datos cargados previamente, mediante el uso del módulo 'pd.DataFrame.merge'

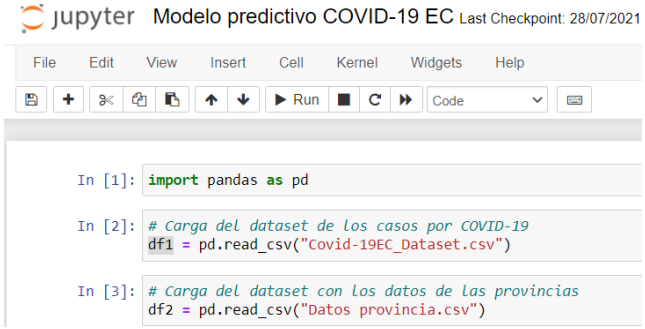


Figura 3. Fase de Selección de datos.

definiendo los datos del campo "provincia" como indicador común y se lo almacenó en la variable "df1". Tras la ejecución de este módulo se obtuvo un nuevo marcos de datos el mismo que contiene los campos de ambos conjuntos de datos (ver Figura 7).

Para la siguiente etapa se dio paso a las tareas de limpieza y depuración de datos. Por ello se empleó el módulo '.info()' el cual proporciona la información de los campos del conjunto de datos

e-ISNN 2550-6730 | Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022 | DOI: 10.33936/isrtic.v6i1.4366

```
Fecha,día,mes,año,casosConfirmados,fallecimientos,provincia,región
30/04/2020,30,4,2020,187,19,Esmeraldas,Costa
29/04/2020,29,4,2020,173,19,Esmeraldas,Costa
28/04/2020,28,4,2020,164,17,Esmeraldas,Costa
27/04/2020,27,4,2020,143,15,Esmeraldas,Costa
23/04/2020,23,4,2020,115,12,Esmeraldas,Costa
22/04/2020,22,4,2020,104,7,Esmeraldas,Costa
21/04/2020,21,4,2020,102,7,Esmeraldas,Costa
20/04/2020,20,4,2020,102,7,Esmeraldas,Costa
19/04/2020,19,4,2020,89,7,Esmeraldas,Costa
18/04/2020,18,4,2020,86,7,Esmeraldas,Costa
17/04/2020,17,4,2020,85,7,Esmeraldas,Costa
16/04/2020,16,4,2020,53,7,Esmeraldas,Costa
15/04/2020,15,4,2020,44,7,Esmeraldas,Costa
14/04/2020,14,4,2020,41,7,Esmeraldas,Costa
13/04/2020,13,4,2020,38,7,Esmeraldas,Costa
12/04/2020,12,4,2020,38,7,Esmeraldas,Costa
11/04/2020,11,4,2020,31,7,Esmeraldas,Costa
10/04/2020,10,4,2020,31,7,Esmeraldas,Costa
```

Figura 4. Estructura del conjunto de datos original.

```
In [7]: # Visualizamos el dataset con los datos de las provincias
        df2
```

Out[7]:

	provincia	región	población	superficieKm2	densidadPoblacion
0	Azuay	Sierra	881394	8309.58	106.069521
1	Bolivar	Sierra	209933	3645.38	53.209830
2	Cañar	Sierra	281396	3146.08	89.443371
3	Carchi	Sierra	186869	3700.45	49.430359
4	Chimborazo	Sierra	524004	6499.72	80.619473
5	Cotopaxi	Sierra	488716	6108.23	80.009430
6	El Oro	Costa	715751	5766.88	124.119363
7	Esmeraldas	Costa	643654	16132.23	39.898138
8	Galápagos	Insular	33042	8010.00	4.125094
9	Guayas	Costa	4387434	15430.40	284.337023
10	Imbabura	Sierra	476257	4587.51	103.816013
11	Loja	Sierra	521154	11062.73	47.109867
12	Los Rios	Costa	921763	7205.27	127.929002
13	Manabí	Costa	1562079	18939.60	82.476874
14	Morona Santiago	Amazonia	196535	24059.40	8.168741
15	Napo	Amazonia	133705	12542.50	10.660155
16	Orellana	Amazonia	161338	21692.10	7.437839
17	Pastaza	Amazonia	114202	29641.37	3.852791
18	Pichincha	Sierra	3228233	9535.91	338.534340
19	Santa Elena	Costa	401178	3690.17	108.715217
20	Santo Domingo	Costa	458580	3446.65	133.050934
21	Sucumbios	Amazonia	230503	18084.42	12.745844

Figura 5. Previsualización de “Covid-19EC_dataset.csv”.

```
In [10]: # Concatenamiento de los dataset
         df1 = pd.DataFrame.merge(df1, df2, how="left", on="provincia")
         df1
```

Out[10]:

	fecha	día	mes	año	casosConfirmados	casosDiarios	fallecimientos	nuevosFallecimientos	provincia	región_x	región_y	población	superficieKm2
0	2020-03-13	13	3	2020	0	0	0	0	Azuay	Sierra	Sierra	881394	8309.58
1	2020-03-14	14	3	2020	1	1	0	0	Azuay	Sierra	Sierra	881394	8309.58
2	2020-03-15	15	3	2020	1	0	0	0	Azuay	Sierra	Sierra	881394	8309.58
3	2020-03-16	16	3	2020	1	0	0	0	Azuay	Sierra	Sierra	881394	8309.58
4	2020-03-17	17	3	2020	5	4	0	0	Azuay	Sierra	Sierra	881394	8309.58
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2460	2020-06-26	26	6	2020	466	0	16	0	Zamora Chichipe	Amazonia	Amazonia	120416	10548.28
2461	2020-06-27	27	6	2020	496	30	16	0	Zamora Chichipe	Amazonia	Amazonia	120416	10548.28
2462	2020-06-28	28	6	2020	523	27	16	0	Zamora Chichipe	Amazonia	Amazonia	120416	10548.28
2463	2020-06-29	29	6	2020	523	0	16	0	Zamora Chichipe	Amazonia	Amazonia	120416	10548.28
2464	2020-06-30	30	6	2020	536	13	16	0	Zamora Chichipe	Amazonia	Amazonia	120416	10548.28

2465 rows × 14 columns

Figura 6. Previsualización de “Datos provincia.csv” cargados a Pandas mediante el módulo ‘pd.read_csv’.

y su estructura. Con ello se pudo observar que el nuevo marcos de datos está estructurado con 14 columnas cada una de ellas con # de entradas lo cual descarta la existencia de campos vacíos. Además de esto se observó que la columna del campo “fecha” está definida como ‘object’, lo cual representaría inconvenientes al momento de generar series temporales ya que es necesaria la presencia de la variable de tipo ‘date’.

```
In [4]: # Visualizamos el dataset de los casos por COVID-19
        df1
```

Out[4]:

	fecha	día	mes	año	casosConfirmados	casosDiarios	fallecimientos	nuevosFallecimientos	provincia	región
0	2020-03-13	13	3	2020	0	0	0	0	Azuay	Sierra
1	2020-03-14	14	3	2020	1	1	0	0	Azuay	Sierra
2	2020-03-15	15	3	2020	1	0	0	0	Azuay	Sierra
3	2020-03-16	16	3	2020	1	0	0	0	Azuay	Sierra
4	2020-03-17	17	3	2020	5	4	0	0	Azuay	Sierra
...	...	...	...	...	...	...	...	...	...	...
2460	2020-06-26	26	6	2020	466	0	16	0	Zamora Chichipe	Amazonia
2461	2020-06-27	27	6	2020	496	30	16	0	Zamora Chichipe	Amazonia
2462	2020-06-28	28	6	2020	523	27	16	0	Zamora Chichipe	Amazonia
2463	2020-06-29	29	6	2020	523	0	16	0	Zamora Chichipe	Amazonia
2464	2020-06-30	30	6	2020	536	13	16	0	Zamora Chichipe	Amazonia

2465 rows × 10 columns

Figura 7. Concatenación de los conjuntos de datos “Covid-19EC_dataset.csv” y “Datos provincia.csv.

2.4.3. Transformación de datos

Esta etapa de la metodología se realizaron las modificaciones necesarias al conjunto de datos con la finalidad proveer una estructura de los datos adecuadas para el modelo propuesto. Se identificó un defecto en el tipo de variable del campo “fecha”, fue necesario transformar dicha variable a tipo ‘date’ con la ayuda del módulo ‘pd.to_datetime’ como se observa en la Figura 8.

Se realizó la conversión de la columna ‘fecha’ de ser una columna que contiene una variable a ser un indicador de las demás columnas, para ello se empleó el módulo ‘set_index()’ sobre el campo mencionado. Este proceso tuvo lugar con el fin de que la estructura del conjunto de datos permita generar un modelo con series temporales más adelante.

2.4.4. Minería de Datos

Concluidas las fases anteriores se obtuvo una base de datos lista para continuar el proceso de minería de los datos, el mismo que se comprendió de tres etapas, partiendo desde la búsqueda de patrones entre el conjunto de datos, los mismos que permitan identificar una correlación entre ellos. Para cumplir esta tarea se empleó la librería Seaborn y Matplotlib de Python.

Se identificaron tres correlaciones importantes siendo estas: casos confirados-fallecimientos, casos confirmados-población y casos confirmados-densidad de la población, las mismas que poseen un grado de correlación equivante a 0.91, 0.76 y 0.61 correspondientemente. Ver Figura 9.

Después de analizar los primeros resultados se identificó que el numero de los casos confirmados (“casosConfirmados”)

e-ISNN 2550-6730 | **Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022** | **DOI: 10.33936/isrtic.v6i1.4366**

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

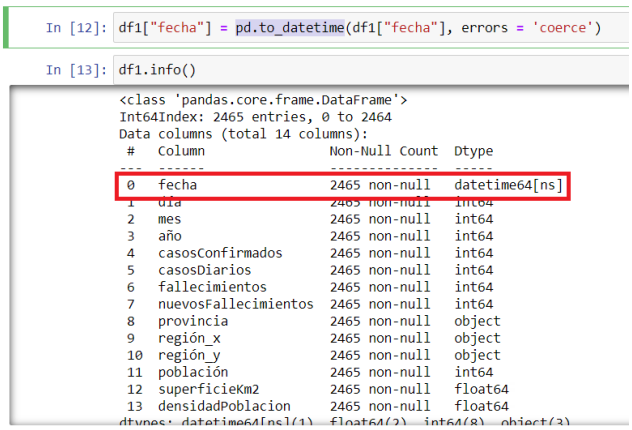


Figura 8. Transformacion Variable 'fecha' de tipo ''object' a tipo 'datetime'.

Figura 9. Gráfica de calor con los índices de correlación.

es la variable que se encuentra presente en cada una de las correlaciones (Ver Figura 10), ademas de ser esta el objetivo a predecir de este modelo. Por esta razón, se realizó un segundo analisis comparando la dependencia de una variable a otra, para realizar esta tarea se empleo el módulo 'sns.pairplot' de la librería Matplotlib.

Del análisis de los resultados, se identificó que a mayor número de habitantes y un indice más elevado de la densidad población en una provincia mayor resulta el incremento de los casos confirmados por COVID-19 y esto representa tambien una elevación en la tasa de fallecimiento por el virus. En la siguiente sección, se relatan los resultados obtenidos a partir de la aplicación del modelo seleccionado.

3. Resultados y Discusión

Posterior al análisis e interpretación de toda la información, se prosigió a buscar un modelo que genere las predicciones en razón a las variables previamente identificadas. Para ello, se consideró trabajar primero sobre un número reducido de datos a traves del módulo '.sample()' que permitió tomar una muestra, siendo esta el diez por ciento de los registros del conjunto de datos. A continuación, se ha se describen los experimentos con algoritmos de predicción para indentificar cual de ellos se adapta mejor al modelo iniciando por los algoritmos de regresión.

3.1. Primer experimento (regresión lineal)

Condiderando los resultados obtenidos de la matriz de correlación, decidió realizar pruebas con modelos de regresión lineal y polinolial, para ello se seleccionaron las variables con mayor correlación siendo estas la 'densidad poblacional' y los 'casos confirmados'.

Para este modelo se utilzó la librería Sckit-learn y el módulo 'LinearRegression', con el fin de encontrar la tendían de la variable 'casos confirmados' con dependincia en la variable 'densidad poblacional', expresada en habitantes por metro cuadrado. Para calcular el grado de presición del modelo se utilizó el error medio cuadrático como métrica de rendimiento, siendo esta 0.38 como podemos observar en la Figura 11.

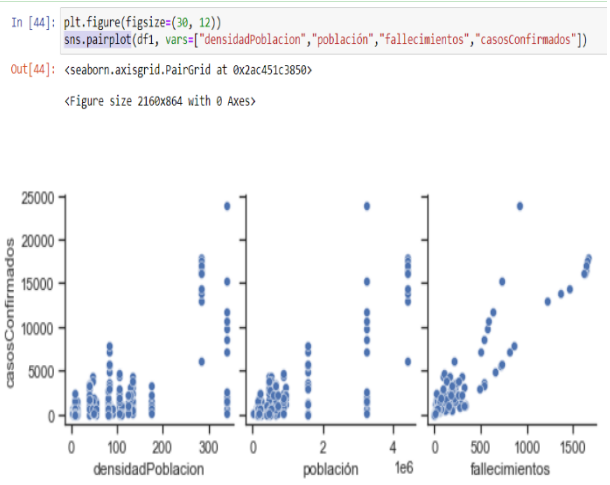
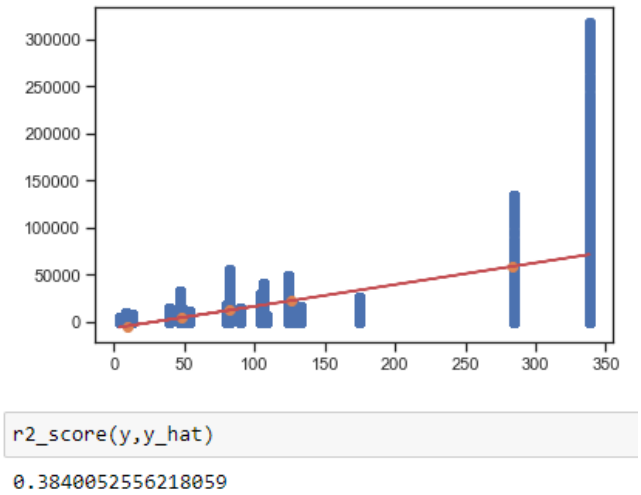


Figura 10. Dependencia de la variable "casosConfirmados" con respecto a las variables "desnsidadPoblacion", "población" y "fallecimientos".

e-ISNN 2550-6730 | Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022 | DOI: 10.33936/isrtic.v6i1.4366




```
r2_score(y,y_hat)
```

```
0.38400525562180 59
```

Figura 11. Primer experimento, definición de campos para regresión lineal.

Debido a los resultados de la métrica de rendimiento, se procedió a ejecutar el modelo de regresión polinomial, escalando con polinomios del grado 1 al 10 con el fin de encontrar el mejor ajuste para el modelo comparando cada uno de ellos. Siendo el polinomio de tercer grado el que presentó una métrica más alta con 0.54 de rendimeinto en compararion a los demás polinomios como se puede observar en la Figura 12.

Posteriormente se procedió a generar las predicciones de los casos confirmados con relación a la cantidad de habitantes por metro cuadrado, a partir del modelo de regresión polinomial con polinomio de tercer grado. Se obtuvieron pronósticos de los casos confirmados en relación con la cantidad de habitantes por metros cuadrados, como se observa en la Figura 13.

Para avaluar los resultados del modelo se tomaron como muestra las provincias con una densidad poblacional aproximada y se compararon los casos confirmados del pronostico con los datos reales con fecha 28 de marzo del 2022. Los mismo que se detallan en la Tabla 1.

Llegado a este punto se implementó el modelo secuencial de memoria de largo plazo utilizando la librería Keras con una capa de entrada, dos capas oculta y una capa de salida, donde las capas ocultas son capas LSTM, ademas la primera de ella pose 4 unidades ocultas y la segunda solo posé 1 unidada. Posterior a ello se evaluó el modelo mediante un historial de perdida, el cual se configuró para el calculo de dos mil iteracicones parada uno de los datos del conjunto de entrenamiento equivalente al 70% del total de los datos, tal como se observa en la ilustración 37. Una vez procesado los datos se pudo obersar que el nivel perdida del modelo era satisfactorio, como se observa en la ilustación 38, ya que una menor perdida proporciona una mejor predicción.

Una vez evaluado el modelo y obtenido un índice de perdida bajo, se procede a realizar las predicciones, basadas en los casos confirmados. Para ello se tomó de muestra las provincias de: Bolívar, Cañar, Loja y Pichincha como se observa en la Figura 14.

Las predicciones se realizaron a partir de los registros de casos confirmados por COVID 19 durante el primer trimestre del año 2022, con extensión hasta el 28 marzo. Se obtuvo el pronostico de 7 días en el futuro (04/04/2022) para las provincias del conjunto de muestra. Se observa que el modelo secuencial LSTM proporciona una precisión del %, para predecir la tendencia de casos confirmados por COVID 19, el desgloce de los resultados se puede observar en la Tabla 2.

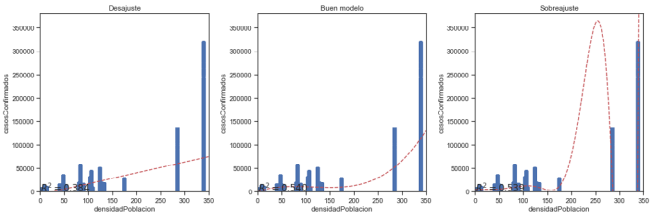


Figura 12. Regresión Polinomial. Ajuste polinomio grado (1-10).

Este trabajo de investigación permitió analizar la información respecto a la crisis sanitaria por Covid-19 y predecir la incidencia de esta pandemia en el Ecuador mediante el uso de procesos y metodologías de minería de datos que se describen en las secciones anteriores. Esta aplicación se constituye en una

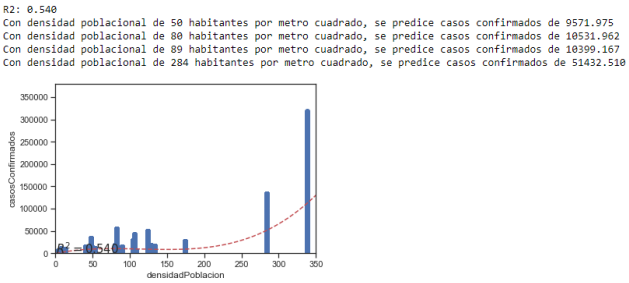


Figura 13. Regresión Polinomial, pronóstico de los casos confinados en relación con la densidad poblacional.

herramienta que aporta a mejor toma de decisiones para prevenir y precautelar la salud de los ecuatorianos.

Tras el análisis correlacional entre los datos de las provincias del Ecuador y los casos confirmados por COVID-19 en cada una de ellas, se identificó que, para los casos particulares entre las provincias de Pichincha y Guayas, el factor más influyente no fuel el total de la población, si no la cantidad de habitantes por metros cuadrado dentro de la provincia. Para el caso de las Galápagos, ninguno de los factores mencionados anteriormente desempeño un papel crucial en la propagación del virus, ya que se observó que los casos nuevos por contagios de COVID 19 se prolongaron en la región y dentro algunos periodos fueron nulos. Se presume que los factores más influyentes fueron las medidas de bioseguridad rigurosas que se venían tomando incluso antes de iniciar la pandemia por mediadas de conservación de las especies endémicas de la región. Al no contar con un indicar que pueda medir este factor no fue posible realizar el estudio de este dentro del impacto de los contagios. Los resultados obtenidos en

e-ISNN 2550-6730 | **Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022** | **DOI: 10.33936/isrtic.v6i1.4366**

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

Tabla 1. Resultados del modelo predictivo: Regresión polinomial.
Fuente: Los autores.

Provincia	Datos reales 23/03/2022		Predicción: Regresión polinomial		Precisión
	Hab. m^2	Casos conf.	Hab. m^2	Casos conf.	
Carchi	49.43	13.111	50	9.572	73 %
Chimborazo	80.62	15.866	80	10.532	66 %
Cotopaxi	80.01	17.899	80	10.532	59%
Cañar	89.44	14.335	89	10.399	73%
Guayas	284.34	134.054	284	51.433	38%

ambos experimentos nos permiten concluir que las herramientas tecnológicas y, en este caso con la aplicación de ciencias de datos cumplen un papel importante en el aporte científico para el desarrollo de soluciones importantes para el país y el mundo.

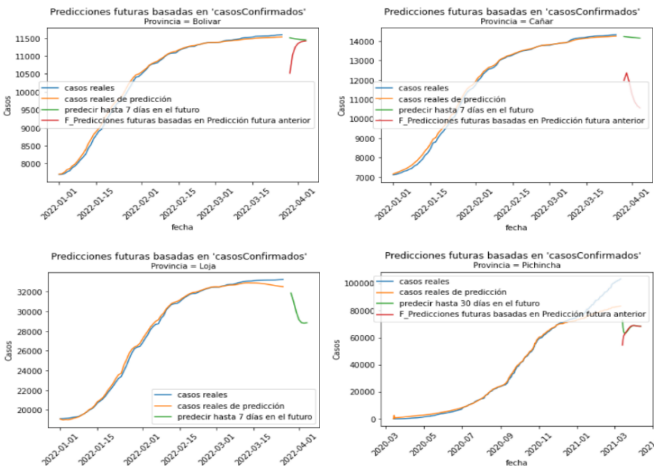


Figura 14. Predicciones conjunto de muestra.

Tabla 2. Resultados del modelo secuencial LSTM.
Fuente: Los autores.

Provincia	Casos confirmasdos (04/04/2022)	Predic-cion	Diferencia	Porcentaje precisión
Bolivar	11640	11447	193	98.34%
Cañar	14388	14160	228	98.42%
Loja	33357	31896	1461	95.62%
Pichincha	319676	298791	20885	93.47%
PROMEDIO PRECISIÓN DEL MODELO				**96.46%**

4. Conclusiones

El estado del arte en técnicas de minería de datos y modelos predictivos fue estudiado y revisado, con el fin de explorar y comprender la manera en que estas técnicas han sido aplicadas en respuesta a la emergencia sanitaria que se presentó por el Covid-19. Para ello, en la investigación fue documentada la metodología KDD junto a las técnicas predictivas que se aplicaron en este proyecto.

Se preparó un conjunto de datos a partir de los datos de Covid-19, emitidos por el Ministerio de Salud Púlica del Ecuador mediante sus boletines epedemiológicos, con un periodo desde el 13 de marzo del 2020 al 28 de marzo del 2022. Con el fin de contar con un conjunto de datos que se puedan procesar e implementar en modelos predictivos. Las limitaciones se dieron debido a la irregularidad de los datos en los registros oficiales, obligó a emplementar técnicas que ayudaron a detectar las anomalías y que estas sean depuradas en el producto final.

Se aplicaron algoritmos de minería de datos de regresión y redes neurales recurrentes para implementarlos en los modelos predictivos de regresión polinomial y memorias de corto y largo plazo respectivamente. Entre estos modelos se pudo observar que las memorias de corto y largo plazo generan mejores pronósticos de casos confirmados por Covid-19, alcanzando una presición general de hasta el 96.46%. Como trabajos futuros se propone actualizar los datos para realizar predicciones futuras incrementando las variables que se presenten en la realidad futura. Además de ampliar los experimentos a más provincias y la mejora contínua de los modelos aplicados.

Agradecimientos

El trabajo de los autores es parcialmente soportado por la Universidad Laica Eloy Alfaro de Manabí.

Contribución de los autores

Juan Carlos Zambrano: Conceptualización, Metodología, Análisis formal. **Patricia Quiroz-Palma:** Redacción – borrador original del artículo, Metodología, Análisis formal. **Alex Santamaría-Philco:** Metodología, Revisión y edición del artículo. **William Zamora:** Revisión y edición del artículo.

Conflictos de interés

Los autores declaran no tener ningún conflicto de interés.

e-ISNN 2550-6730 | Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022 | DOI: 10.33936/isrtic.v6i1.4366

Anexos

Análisis e interpretación de las encuestas aplicadas

Pregunta 1.- ¿Pertenece o perteneció usted al personal de primera línea (médicos, enfermeras, paramédicos, etc....) frente al COVID-19?

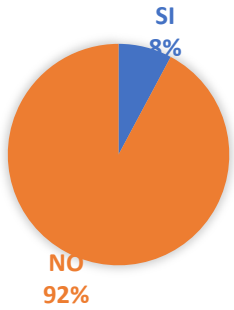


Figura A1. Grafico pregunta 1.

Análisis:

De un total de 128 encuestados que representan al 100%, el 92% que corresponde a 118 encuestados han respondido que no pertenecen al personal de primera línea, y el 8% que pertenece a 10 encuestados han respondido que si pertenecen o pertenecieron al personal que labora o laboró en primera línea en la pandemia de COVID-19.

Interpretación:

La mayor parte de los encuestados pertenecen al grupo de la comunidad estudiantil con conocimientos de campo respecto a la investigación realizada, mientras que la parte restante son profesionales de primera línea en la pandemia de COVID-19.

Pregunta 2.- ¿Conoce usted que es la Minería de Datos?

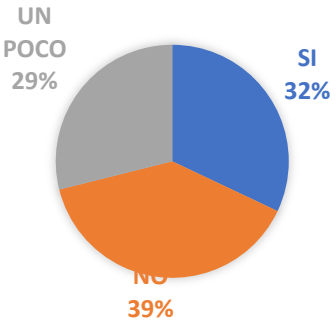


Figura A2. Gráfico pregunta 2.

Análisis:

De un total de 128 encuestados que representan al 100%, el 39% que corresponde a 50 encuestados han respondido que no conocen que es la minería de datos, y el 32% que pertenece a 41 encuestados han respondido que si conocen y el 29% que corresponde a 37 encuestados respondieron que tienen un conocimiento limitado del tema.

Interpretación:

Mas de la mitad de los encuestados tienen la noción de lo que es la minería de datos sus tendencias y alcances a la hora de resolver problemas en diversos campos de acción.

Pregunta 3.- ¿Sabía usted que en el área de la salud la Minería de Datos puede tener un valor científico o investigador que ayude a determinar causas de determinadas patologías o a identificar poblaciones de riesgo?

Figura A3. Gráfico pregunta 3.

Análisis:

De un total de 128 encuestados que representan al 100%, el 54% que corresponde a 69 encuestados han respondido que sí estaban familiarizados con aplicación de la minería de datos en estas temáticas, y el 46% que pertenece a 59 encuestados han respondido que no estaban informados al respecto.

Interpretación:

Un poco más de la mitad de los encuestados están al tanto de los avances de la minería de datos en las investigaciones de causas patológicas, esto es un indicador de que la mayor parte de los encuestados tienen la noción de que se presente con lograr con la esta investigación.

Pregunta 4.- ¿Cree usted que se esté aplicando Minería de Datos para estudiar el COVID-19 en Ecuador?

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

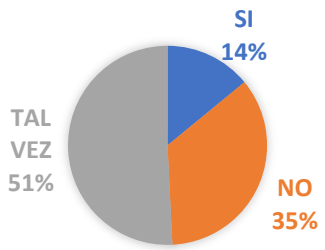


Figura A4. Gráfico pregunta 4.

Análisis:

De un total de 128 encuestados que representan al 100%, el 51% que corresponde a 65 encuestados han respondidos que no están seguros de si se aplica esta tecnología en el país, el 36% que pertenece a 45 encuestados han respondido que no se está aplicando esta mientras que el 14% restante y que corresponde a 18 encuestados consideran que si se está aplicando minería de datos para estudiar el COVID-19 en el país.

Interpretación:

La mitad de los encuestados desconocen sí se está aplicando minería de datos para estudiar la pandemia en el país, la mitad restante se reparte entre grupos que afirman si se aplica o no minería de datos como herramienta de estudio.

Pregunta 5.- Si se empleara Minería de Datos y algoritmos para predecir la incidencia del COVID-19 en Ecuador. ¿Cuál considera usted que sería su impacto?

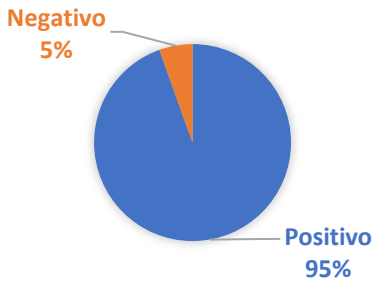


Figura A5. Gráfico pregunta 5.
Fuente: Encuestas-Elaboración propia

Análisis:

De un total de 128 encuestados que representan al 100%, el 95% que corresponde a 121 encuestados han respondido que una vez que se logre predecir la incidencia del COVID-19 esta investigación tendría un impacto positivo, y el 5% que pertenece a 7 encuestados han respondido que no tendrá un buen impacto.

Interpretación:

Con las respuestas obtenidas en esta sección de la encuesta se logra determinar el índice de aceptación de esta investigación, como un apoyo a los aconteceres que si ven en el país a casusa de la pandemia.

Pregunta 6.- Observando su propio entorno. ¿Qué factores considera usted que están relacionados con la incidencia del COVID-19 en su localidad?

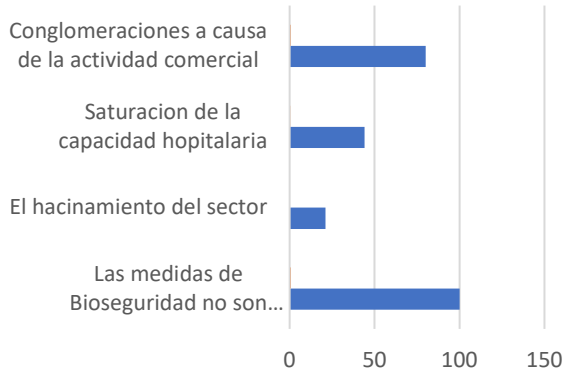


Figura A6. Gráfico pregunta 6.

Análisis:

De un total de 128 encuestados que representan al 100%, en esta serie de pregunta de opciones múltiples el 78% que corresponde a 100 encuestados han respondido que el mal cumplimiento de las medidas de bioseguridad es el factor que tiene una relación más estrecha con la incidencia del COVID-19 en el País, el 62% que corresponde a 80 encuestados señalaron también que las conglomeraciones a causa de actividades de carácter económico son el segundo factor responsable de la incidencia investigada, el 34% que corresponden a 44 encuestas señalan la saturación en los hospitales como otro factor de relevancia mientras que el 16% que equivale a 21 de las encuestas repartidas indican que el hacinamiento de la población es otro de los factores que participa con un menor grado de efectividad en la incidencia del virus en el país.

e-ISNN 2550-6730 | Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022 | DOI: 10.33936/isrtic.v6i1.4366

Interpretación:

La escasa ejecución de las medidas de bioseguridad sanitaria sería el factor con mayor relación en las incidencias del COVID-19 presentadas en el país según las encuestas, a esto se le suman los demás factores que tienen participación en este acontecimiento.

Pregunta 7.- Una vez que se haya implementado la minería de Datos y se logre anticipar la incidencia del COVID-19. ¿De qué manera considera usted que beneficiaría esto al Ecuador?

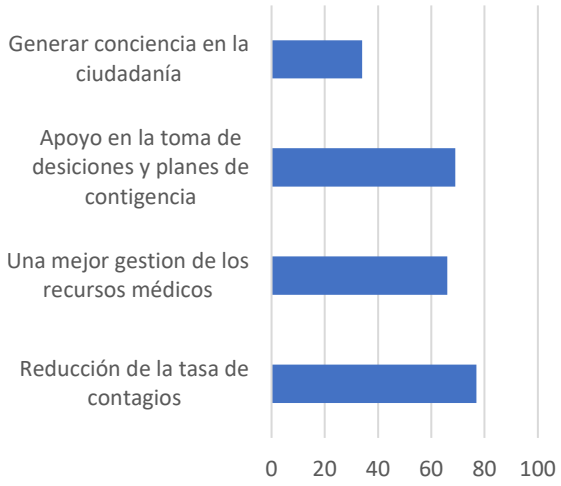


Figura A7. Gráfico pregunta 7.

Análisis:

De un total de 128 encuestados que representan al 100%, en esta serie de pregunta de opciones múltiples el 60.2% que corresponde a 77 encuestas responden a que mayor beneficio una vez implementada la propuesta es la reducción en la tasa de contagios por COVID-19, el 53.9% que corresponde a 69 de las encuestas señalan también que implementación de esta propuesta apoyara a las autoridades en la toma de decisiones, el 51.6% que corresponden a 66 encuestas señalan que mejorará la gestión de los recursos médicos el 26.6% que equivale a 34 indican que predecir la incidencia ayudaría a generar conciencia en la ciudadanía

Interpretación:

La mayoría de los encuestados coinciden en que la aplicación de este estudio jugaría un papel en la reducción de la tasa de contagios con COVID-19, gran parte de los encuestados señalan también predecir la incidencia del virus apoyaría a las autoridades a tomar mejores decisiones en esta lucha contra la pandemia.

Pregunta 8.- ¿Qué tan necesario cree usted que se debe aplicar Minería de Datos para predecir la incidencia del COVID-19 en el País?

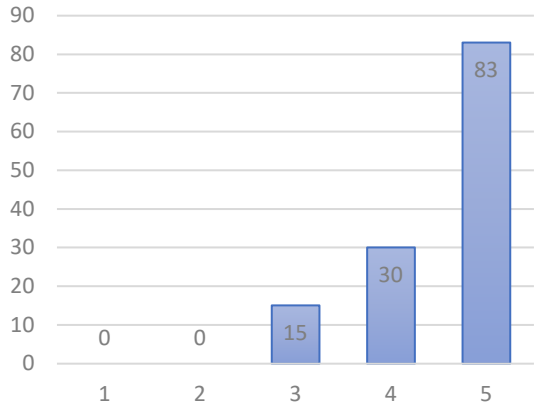


Figura A8. Gráfico pregunta 8.

Análisis:

De un total de 128 encuestados que representan al 100%, el 64.8% que corresponde a 83 de las encuestas indican con el mayor nivel de necesidad la implementación minería de datos para predecir la incidencia del COVID-19 en el País, el 23% que equivale a 30 de las encuestas reflejan un grado de necesidad de nivel 4, mientras que el 11% restante que señala que no están necesario aplicar Minería de datos para la predecir la incidencia del virus.

Interpretación:

En esta sección de la encuesta mediante el empleo de preguntas con respuestas de escala lineal donde 1 significa el más bajo nivel de necesidad y 5 el nivel de necesidad más alto destaca que la mayoría de los encuestados coinciden en que se debe aplicar minería de datos para predecir la incidencia del COVID-19 en el país.

Referencias bibliográficas

Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Kalhori, S. R. N. (2020). Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. *JMIR public health and surveillance, 6*(2), e18828.

Celestine, I., Ali, K. B., Atharva, P., & R, S. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health*, 357.

Ray, D., Salvatore, M., Bhattacharyya, R., Wang, L., Du, J., Mohammed, S., ... & Mukherjee, B. (2020). Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: data science call to arms. *Harvard data science review, 176*(3), 139-148.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI*

e-ISNN 2550-6730 | **Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022** | **DOI: 10.33936/isrtic.v6i1.4366**

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

magazine, 17(3), 37-37.

Wang, C. J., Ng, C. Y., & Brook, R. H. (2020). Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. *Jama, 323*(14), 1341-1342.

MSP. (2022). Vacunómetro. Obtenido de

https://app.powerbi.com/view?

r=eyJrIjoiYTkzNTFkMmUtZmUzNi00NDcwLTg0MDEtNjFk

NzhhZTg5ZWYyIiwidCI6IjcwNjIyMGRiLTliMjktNGU5MS

1hODI1LTI1NmIwNmQyNjlmMyJ9&pageName

=ReportSection

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses. *Journal of advanced research, 24,* 91-98.

Salcedo, F., & Salcedo, G. (2021). Modelos predictivos de los contagios de la COVID-19 para la provincia de Loja-Ecuador. *Revista Digital Novasinergia, 4*(2), 62-77

e-ISNN 2550-6730 | Vol. 6, Núm. 1 (12-23): Enero-Junio, 2022 | DOI: 10.33936/isrtic.v6i1.4366